Reverse engineering of transcriptional and regulatory networks using gene expression

Alexander Lachmann
Transcript Abundance Quantification

Transcript abundance estimated through RNASeq can access a higher level of complexity of the Transcriptome, specifically:

- **Gene abundance**
- **Isoform abundance** (alternative splicing)
- **Aberrations** (fusion genes)
- **Allele abundance**
- **RNA editing**
- **Novel transcripts**

Moreover, it can do so without any prior knowledge of the organism(s) under investigation.
Microarrays vs. RNAseq

Signal vs. Quantity plots

Microarray signal

Real quantity of transcript

Saturation

Limit of detection

Microarrays vs. RNASeq
Microarrays vs. RNASEq

Gene length bias

Tarazona 2011
**Microarrays vs. RNAsSeq**

Figure 2. Quantifying expression levels: RNA-Seq and microarray compared

Expression levels are shown, as measured by RNA-Seq and tiling arrays, for *Saccharomyces cerevisiae* cells grown in nutrient-rich media. The two methods agree fairly well for genes with medium levels of expression (middle), but correlation is very low for genes with either low or high expression levels. The tiling array data used in this figure is taken from REF. 2, and the RNA-Seq data is taken from REF. 18.

*Wang et al., 2008*
Allele-Specific Expression

Allele expression indistinguishable within an homozygous organism:

In the crossing, if variations exist between the two alleles (e.g. SNPs and indels in the mRNA) we can measure the alleles separately:

Federico Giorgi

Proximal enhancer
Allele-Specific Expression

First case: **homozygous genotypes** (e.g. Crossing in-bred maize lines)

Allele

<table>
<thead>
<tr>
<th>Father</th>
<th>Mother</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a</td>
</tr>
<tr>
<td>A</td>
<td>a</td>
</tr>
</tbody>
</table>

Gene

Allele expression indistinguishable within an homozygous organism
**Allele-Specific Expression**

First case: **homozygous genotypes** (e.g. Crossing in-bred maize lines)

**Allele**

Father: \( A \) \( A \)  
Mother: \( a \) \( a \)

**Gene**

Allele expression indistinguishable within an homozygous organism

**Father**  
\( A \downarrow \)

**Mother**  
\( a \)

**Crossing**

IF the two alleles are distinguishable, e.g. because they carry enough specific SNPs or indels, the expression of specific alleles can be evaluated.

Is the maternal or the paternal one prevailing?  
Is it because of the allele itself?  
Is it because of a proximal specific enhancer? (Need genomic reads for **phase inferring**)

\( \text{C} \rightarrow \text{T} \)
\( \text{T} \rightarrow \text{G} \)
\( \text{T} \rightarrow \text{G} \)
Allele-Specific Expression

Second case: **heterozygous genotypes** (e.g. Comparing *Vitis vinifera* varieties)

- **Variety 1**: A a
- **Variety 2**: a a
- **Variety 3**: A α
- **Variety 4**: α a

Allele expression can be distinguishable within an heterozygous organism

Are some alleles always prevailing over the others?
Is the proximity of some genomic regions determining a constitutive prevalence of one allele?
Allele-Specific Expression

Alleles can be differentially expressed by chance, or if the read coverage is low. A binomial test can give the p-value of an Allele Ratio (e.g. roughly: the chance of it to be generated by chance).

Other ASE issues:
• Distinguish between close *paralogs* and *alleles*
• Biases in sequencing and read mapping: need to normalize transcriptomic allele ratio with *genomic* allele ratio!
Coexpression

Transcriptional Network

Escherichia coli
Gama-Castro et al., 2008

RT-PCR

RNASeq

Expression measurements
Coexpression

Gene Network
Reverse Engineering

RT-PCR
RNASeq

Expression measurements
Co-regulation and Co-expression

**Common cause**
e.g. Cellulose Synthase (CESA) complex

**Cause-Effect**
e.g. UDP-L-Rhamnose synthesis

- Gene A
- Gene B
- Gene C

- CESA1
- CESA3
- CESA2/5/6/9

- AP2
- GL2
- RHM2
- TF
- enzyme
An example of coexpressed genes

Cellulose Synthase complex

Log2 tRMA normalized expression

Samples

AT4G32410 (CESA1)
AT5G64740 (CESA6)
Transcriptome-wide reverse engineering

Evaluate all gene pairs associations

Expression-based reverse-engineering methods

Coexpression network quality assessment
General methods

1. Correlation  
   (Pearson, Spearman, ...)
2. Mutual Information
3. Linear Regression
4. ...

Conditional methods

1. Partial Correlation
2. Partial Mutual Information
3. LASSO regression
4. ...

Zampieri et al., 2008
Correlation

Pearson Correlation: operates on real values

Spearman Correlation: operates on rank-transformed values

\[ r_{xy} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\text{var}(y)}} \]

Perfect negative correlation

No correlation

Perfect positive correlation

\[ r = 0.75 \]
Partial Correlation

Standard Correlation (zeroth order)
\[ r_{xy} = \frac{cov(xy)}{\sqrt{var(x)var(y)}} \]

Partial Correlation (first order)
\[ r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} \]
Partial Correlation

Standard Correlation (zeroth order)
\[ r_{xy} = \frac{cov(xy)}{\sqrt{\text{var}(x)\text{var}(y)}} \]

Partial Correlation (first order)
\[ r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} \]
Mutual Information

Can detect for any relationship (also *nonlinear*)

\[
H(X) = \sum_{i=1}^{n} p(x_i) \log_b p(x_i),
\]

\[
I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)
\]

Can be *partial* (conditional method)
Mutual Information

Developmental dataset (Atge0100)

Pearson = 0.15
M.I. = 1.04

Mutual information

Pearson correlation

AT4G21860 (Methionine Sulfoxide Reductase 2)

AT1G35340 (ATP-dep protease)
ARACNe

• Used to predict transcription factor targets.
• The core strength is to eliminate the vast majority of indirect interactions typically inferred by pairwise analysis.

Legend: A transcription factor molecule binds to the DNA at its binding site, and thereby regulates the production of a protein from a gene.
Information Theory

\[ I[\text{Mary}; \text{Jane}] \leq \min \{ I[\text{Mary}; \text{Joe}], I[\text{Joe}; \text{Jane}] \} \]

\[ I[\text{TF}_1; \text{Target}] \leq \min \{ I[\text{TF}_1; \text{TF}_2], I[\text{TF}_2; \text{Target}] \} \]
ARACNe Data Processing Inequality (DPI)
Higher order DPI (hARACNe)
Gene Set Enrichment Analysis
• Virtual Proteomics: Inferring global protein activity profiles by network based analysis of gene expression signatures
ARACNe

a) TRANSCRIPTOMICS: DNA → Transcription → RNA → RNA decay → Translation → Protein degradation → Post-translational modification → Inactive protein → Subcellular localization → Active protein

b) GEP

ARACNe network

R1, R2, R3

MoR analysis

T₁₈, T₃, T₁

Spearman correlation

R1

GES

Distribution density

Induced targets

Repressed targets

NES
Modulator Analysis

I[Marj; Joe | Tony]
Mindy

- Dissect context-specific signaling pathways and combinatorial transcriptional regulation.
- Consider triplets of TF, modulator and target.

Kai Wang et al., 2009
Normalize counts

**Multiplexing:** Illumina trick to save money AND reduce biases between samples

Balanced Blocked Design

Confounded Design

One Illumina HiSeq run: 150Gb (i.e. 20 Gb per lane)

Auer et al., 2010
Plate-Seq

Forest Ray
PLATE-Seq motivation

1 experiment
1 library
$1,600

96 experiments
1 barcoded library
$1,600

Sequencing

mRNA expression

Inferred protein activity
Accuracy on subsampled RNA-seq data

Correlation to 30M reads signatures

Correlation to 30M reads

Depth (mapped reads)

Expression

VIPER

1M

2M
PLATE-Seq Protocol Outline

A

• Lyse cells
• Add bar codes, ERCC spike-ins & adapter-linked oligo(dT) primers

• Reverse transcribe
• Destroy residual primers
• Heat Kill RT

• Pool samples

Remaining Steps Occur in a Single Sample

- Second-strand synthesis with adapter-linked random primers
- PCR enrichment
- Size selection and sequencing
QC Measurements

A. Fraction of reads mapping to rRNA

B. Gene body percentile (5' to 3')

C. Number of genes detected

D. Number of uniquely mapped reads
Citrus / Pancancer

- Identify genomic alterations such as copy number alterations (CNV) and single nucleotide mutations modifying activity of Master Regulators.
End